Focus on Your Target: A Dual Teacher-Student Framework for Domain-adaptive Semantic Segmentation
Appendix

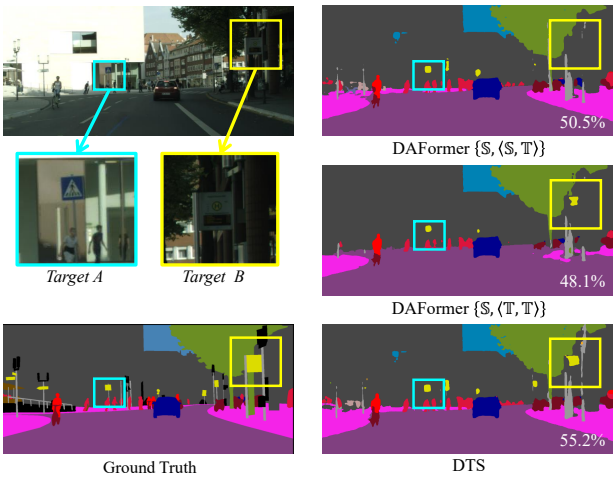

Figure 6. Qualitative and quantitative comparisons of the single teacher-student model trained on different data combination strategies. The numbers shown in the bottom-right corner denote the mIoU of the overall image.

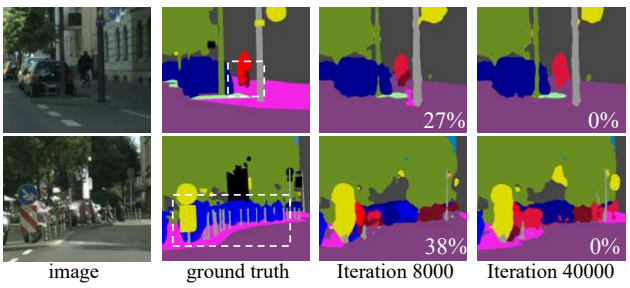

Figure 7. Two more examples of 'conflict' in DAFormer.

A. Examples of the Conflict in Learning

We provide an example of the conflict between the learning and adapting abilities, which we discussed in the *Introduction* section of the main article. This is tested using a single teacher-student framework. We use the same test case as in Figure 1, which has two targets of traffic sign with different appearances. When trained on the original data combination, $\{\mathbb{S}, \langle \mathbb{S}, \mathbb{T} \rangle\}$, the baseline method (DAFormer [1]) shows that *Target B* (an object with a larger transfer difficulty) gradually grows but eventually drops to 0, resulting in an overall mIoU of 50.5%. When the proportion of the target domain is increased using the combination of $\{\mathbb{S}, \langle \mathbb{T}, \mathbb{T} \rangle\}$, the model has a better performance on *Target B* but reports a lower mIoU of 48.1%, as shown in Figure 6. Therefore, improving the adaptability of *Target B* in this framework can potentially deteriorate the learning of other semantic concepts. However, the proposed DTS performs well on *Target B* without harming the recognition of other classes, hence improving the overall mIoU to 55.2%. Except for the example in Figure 1, we show two more samples (bike & motorbike) in Fig 7. DTS improves the IoU in 62% of such cases, demonstrating its generalized ability.

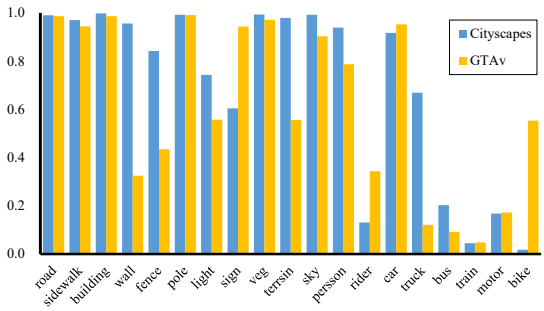

Figure 8. Frequency of subclass occurrence in the source dataset (GTAv) and target dataset (Cityscapes).

B. Statistical Differences between Domains

To make a clear and distinct comparison between the source and target domains, we compute the occurrence frequency of each class in GTAv and Cityscapes. This is done by simply dividing the number of images including a given category by the total number of images – we believe that better metrics can be defined. The comparison of 19 classes is presented in Figure 8. We can see that the traffic sign and truck classes exhibit a large difference between the two datasets, which is consistent with the difficulties during transfer (see the analysis in Section 4.2 of the main article).

Method	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motor	Bike	mIOU
DACS	95.3	67.9	87.5	33.7	30.5	40.2	50.2	56.4	87.7	45.3	87.0	67.4	29.7	89.5	48.0	50.5	2.2	23.0	32.0	53.9
DACS + DTS	96.1	72.2	88.2	38.1	33.6	41.6	52.1	61.4	88.6	50.0	89.1	68.5	38.0	90.6	59.1	58.0	0.2	8.6	14.5	**55.2**
DAFormer	94.7	66.3	87.9	40.7	33.9	37.2	50.2	52.9	87.9	46.5	88.2	69.8	44.2	89.1	43.1	55.8	0.7	24.7	50.7	56.0
DAFormer+DTS	96.4	73.9	88.6	40.0	39.8	42.2	52.2	63.5	88.8	49.6	89.3	70.6	45.6	90.8	61.1	57.0	0.4	33.1	55.4	**59.9**
DACS	81.3	38.9	84.6	15.3	1.7	40.2	45.2	50.3	85.0	–	85.3	70.4	41.9	84.6	–	44.8	–	39.8	57.5	54.2
DACS + DTS	88.7	52.7	85.5	7.2	2.5	40.5	48.9	52.1	86.0	–	87.8	72.5	46.8	83.9	–	43.4	–	46.6	60.9	**56.6**
DAFormer	66.8	29.3	85.0	19.1	2.3	38.7	45.9	51.6	80.8	–	85.9	70.2	41.9	84.9	–	46.0	–	48.9	58.4	53.5
DAFormer+DTS	88.9	52.5	85.1	7.5	2.4	39.7	49.5	52.7	85.6	–	87.0	72.8	47.0	85.0	–	48.0	–	47.7	58.9	**56.9**

Table 11. Segmentation accuracy (IOU, %) of baselines [2, 1] and DTS based on ResNet101 backbone. The top part shows the transfer results for **GTAv→Cityscapes** and the bottom part shows the results for **SYNTHIA→Cityscapes**, where mIoU is computed over 16 classes. All results are averaged over 3 runs.

Options	GTAv	SYNTHIA
$\{\langle\mathbb{T},\mathbb{T}\rangle\}$	69.1	61.1
$\{\langle\mathbb{S},\mathbb{T}\rangle\}$	69.3	59.4
$\{\mathbb{S},\langle\mathbb{T},\mathbb{T}\rangle\}$	70.3	**63.8**
$\{\langle\mathbb{S},\mathbb{T}\rangle,\langle\mathbb{T},\mathbb{T}\rangle\}$	**71.2**	61.2

Table 12. The results of DTS with more data mixing options. We show the mIoU of the 16 classes in SYNTHIA and make this comparison based on the DAFormer baseline.

Except for these two representative classes, we also find that the fence, traffic light, rider, and bike classes are statistically improved by DTS on three strong baselines (see Table 2), despite them having substantial differences between GTAv and Cityscapes.

C. Results on the CNN-based Backbones

We present the class-wise segmentation accuracy of the CNN-based backbones on both the GTAv→Cityscapes and SYNTHIA→Cityscapes benchmarks in Table 11. The proposed DTS approach achieves significant improvements in the traffic sign and truck categories when GTAv is used as the source data, which is consistent with the results obtained using a transformer backbone (see the analysis in the previous section). In the case of using SYNTHIA as the source data, our approach outperforms other methods in the road, sidewalk, person, and rider classes, which are easily confused with each other (road *vs.* sidewalk, person *vs.* rider).

D. More Data Combination Options

We propose two data combinations to tune the proportion of target data in our method: $\{\mathbb{S},\langle\mathbb{T},\mathbb{T}\rangle\}$ and $\{\langle\mathbb{S},\mathbb{T}\rangle,\langle\mathbb{T},\mathbb{T}\rangle\}$. While there are two other options to increase the focus on the target domain as well, namely $\{\langle\mathbb{T},\mathbb{T}\rangle\}$ and $\{\langle\mathbb{S},\mathbb{T}\rangle\}$. The two options only include one type of data and have a disadvantage compared to the previous two combinations (see Table 12). $\{\langle\mathbb{T},\mathbb{T}\rangle\}$ means that the second teacher-student group achieves the ability of **learning** only from the pseudo labels provided by $f_1^{\text{st}}(\mathbf{x};\boldsymbol{\theta}_1^{\text{st}})$. As a result, noise can easily persist without accurate supervision from the source data. $\{\langle\mathbb{S},\mathbb{T}\rangle\}$ has a similar data proportion as $\{\mathbb{S},\langle\mathbb{T},\mathbb{T}\rangle\}$ but cannot solve the problem of inconsistency between the training and testing data. Using $\langle\mathbb{T},\mathbb{T}\rangle$ as the training domain is an effective way to bridge the gap between the training and testing phases. Therefore, the two combinations containing two different types of data used in DTS are more suitable for balancing the **learning** and **adapting** abilities.

References

[1] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[2] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. 2